Extracting regular mobility patterns from sparse CDR data without a priori assumptions

Burkhard, Oliver ; Ahas, Rein ; Saluveer, Erki ; Weibel, Robert

Abstract: In this work we present two methods that can extract habitual movement patterns and reconstruct the underlying movement of users from their call detail records (CDR) in a way that works for users with only moderate numbers of CDRs and that does not make any prior assumptions on the behaviour of the users. The methods allow for a more comprehensive user base in large-scale studies due to the fact that users that might otherwise have to be discarded can also be analysed. The first one is computationally not overly intense and is based on association mining. The second one, which we named DAMOCLES, is based on extracting idiosyncratic daily patterns from clustered daily activities. The methods are evaluated on real data of 140 users over an average of 200 days against benchmarks using assumptions commonly found in the literature such as a work week from Monday to Friday on GPS ground truth. Both methods clearly outperform the benchmarks and for many users retrieve similar regularities. Additionally a simulation study is performed that allows to evaluate the methods in a more controlled environment.

DOI: https://doi.org/10.1080/17489725.2017.1333638

To appear in the *Journal of Location Based Services*
Vol. 00, No. 00, Month 20XX, 1–19

Extracting Regular Mobility Patterns From Sparse CDR Data Without *a priori* Assumptions

Oliver Burkhard[a,†], Rein Ahas[b,‡], Erki Saluveer[c,§], Robert Weibel[a,¶]

[a] *Department of Geography, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland*;
[b] *Department of Geography, University of Tartu, Vanemuise 46, 51014 Tartu, Estonia*;
[c] *Positium Ltd., Õpetaja 9, 51003 Tartu, Estonia*

(*Received 00 Month 20XX; final version received 00 Month 20XX; accepted 00 Month 20XX*)

In this work we present two methods that can extract habitual movement patterns and reconstruct the underlying movement of users from their call detail records (CDR) in a way that works for users with only moderate numbers of CDRs and that does not make any prior assumptions on the behaviour of the users. The methods allow for a more comprehensive user base in large scale studies due to the fact that users that might otherwise have to be discarded can also be analysed. The first one is computationally not overly intense and is based on association mining. The second one, which we named DAMOCLES is based on extracting idiosyncratic daily patterns from clustered daily activities.

The methods are evaluated on real data of 140 users over an average of 200 days against benchmarks using assumptions commonly found in the literature such as a work week from Monday to Friday on GPS ground truth. Both methods clearly outperform the benchmarks and for many users retrieve similar regularities. Additionally a simulation study is performed that allows to evaluate the methods in a more controlled environment.

Keywords: Location Prediction; Mobility; Spatiotemporal Clustering

[†]Corresponding author. Email: oliver.burkhard@geo.uzh.ch
[‡] Email: rein.ahas@ut.ee
[§] Email: erki.saluveer@positium.com
[¶] Email: robert.weibel@geo.uzh.ch

1. Introduction and aim of this work

For city and transport planners it is indispensable to know the transport demand which is determined by origin, destination and the time of day that people move from place to place. In order to estimate the demand, traditionally travel surveys have been conducted by asking participants to fill in travel diaries. More recently, GPS-sensors have been considered (sometimes jointly with diaries) as an alternative to the diaries, as they are not prone to forgetting trips and the cognitive burden on the participants is typically lower (Vij and Shankari 2015; Krumm, Rouhana, and Chang 2015; Shen and Stopher 2014; Houston, Luong, and Boarnet 2014). Both sources usually lead to rich information on the surveyed population but they are typically also rather expensive and can therefore not be carried out frequently.

Call Detail Records (CDR) contain information about calls and text messages users send and receive, as well as about the telephone mast that received/transmitted them. They can be used as an alternative to GPS to infer estimates of the temporally varying locations of significant fractions of the population of a given geographical area, such as a city, region or country (Ahas et al. 2015; Trasarti et al. 2015; Doyle et al. 2014; Blondel, Decuyper, and Krings 2015). Similar to GPS-trajectories, CDR datasets are typically "data rich" but semantically very poor. However, there are differences between GPS-based and CDR-based analyses: GPS provides a relatively accurate and comprehensive record of the mobility of an individual and the semantic poverty vis-à-vis the travel diaries can be remedied at least partially by machine learning techniques (Rinzivillo et al. 2014). CDR come with the benefit of being available for significant parts of a population and hence can *a priori* provide a good sample of the population in regions where cell phone penetration is high. A second important characteristic is the lack of burden on the individuals whose movement is being analysed. Additionally, no recruitment is necessary, as only already available information is analysed.

However, all of these benefits come at the price of spatial imprecision and temporal irregularity and sparsity. The spatial imprecision is mostly given by the fact that only information on the cell tower that routed the call or the message is available. The cells themselves can be of considerable size that may also vary substantially between different towers and connection modes. This makes it very hard to obtain the exact locations of any individual, even at the precise time when there is a CDR. It is technically possible to get estimated locations instead of only cell tower IDs (Widhalm et al. 2015), but this augmented accuracy is typically not available for large samples and long periods of time. The temporal irregularity is not a technological given, but is the result of the calling and texting behaviour of the users. CDR tend to be very unevenly distributed among users but also across time for every user, as shown by Gonzalez, Hidalgo, and Barabási (2008) and in the present work. While it is easy and tempting to discard the users with few CDR or days of users where there is little information – after all there is typically still more than enough data to analyze – we believe that this should not be done as the users discarded based on their calling and texting behaviour may show specific movement behaviour the under-representation of which would bias the estimations. Instead we believe that the methods used should be able to handle users with moderate amount of CDR, even if they can only approximate the underlying behaviour. In the same vein it can be argued that the strength of CDR data – the fact that most people are captured – should not be undermined by making *a priori* assumptions on the hours that people work and on the days that they are off. This would allow for statements about shift and night workers as well as those employees working on weekends that have other days off. The aim of this work is to show and compare

methods that allow getting closer to GPS-like trajectories based on longitudinal CDR data even for the time periods where there is no calling activity and for users that do not produce significant numbers of CDRs. This could potentially broaden the subject base for studies that so far require the installation of a data collection app on a user's phone or handing out GPS sensors. We apply an approach already tested in the GPS context for the problem of finding sequences of activities (Ye et al. 2009) to the problem of trajectory reconstruction and propose an alternative approach based on clustering. We test both of them against ground truth and benchmarks to evaluate their merits. The novelty of the current work is therefore

(1) The transfer of a method from the GPS context using association mining to the context of CDR trajectory reconstruction.
(2) Another method for trajectory reconstruction, called DAMOCLES, to extract regular aspects of mobility patterns of cell phone users based on clustering
    (a) that does not see the CDRs as trajectory that can simply be interpolated but that additionally uses the regularity of the calling and texting behaviour,
    (b) that works for people even if they only have a moderate amount of CDRs and
    (c) that does not make any assumptions on when people work.
(3) A clustering method for cell locations tailored to the bimodal densities of cell tower locations that can be used to infer significant locations of a user.
(4) An evaluation of both reconstruction methods on data collected from 140 people for up to a year
(5) An evaluation of both reconstruction methods on a simulation study that allows for a close control of the parameters

The remainder of the article is structured as follows. Section 2 will summarise the related literature. In Section 3 we present the data. A detailed description of the method is given in Section 4 followed by a presentation of the results in Section 5. The discussion and conclusion can be found in Section 6.

2. Related work

CDR data can be used as a relatively cheap and fast way to obtain information on a large fraction of a population in an area of interest (Steenbruggen, Tranos, and Nijkamp 2015). There are however some challenges that come with this kind of data. As a result of the uniqueness of its advantages it is almost impossible to find matching ground truth at the scale CDR are available and people can often only compare their findings with official statistics at the granularity in which these statistics are available (Calabrese et al. 2013; Janzen et al. 2016) or they make comparisons across different regions (Kung et al. 2014). In addition depending on the exact data source there are obvious concerns of privacy and data may have to be aggregated. The most extreme form of this is when the data is not even available on the individual level but only aggregated on the level of the cells (Louail et al. 2014; Ahas et al. 2015). Even if the information is available on an individual level, if a fine spatial granularity is not the primary interest, aggregating into broader geographical regions can help reduce the uncertainty and noise in the data as well as simplify inference (Tanahashi et al. 2012; Doyle et al. 2014). Depending on the analysis that is to be performed, it can be necessary to restrict the user base by dropping users that do not have enough CDR (Becker et al. 2011; Zhao et al. 2016; Ahas et al. 2010). However, depending on what share of the users is

excluded, it can happen that one sacrifices one of the advantages of CDR: the fact that one can capture a large and arguably representative part of the population. Some methods that achieve very impressive reconstruction results require very dense data, which can come from either users who produce CDR at a rate that substantially surpass averages (e.g. as a result of mobile internet being included in the CDR) (Widhalm et al. 2015) or from users for which not only CDR but also e.g. handover information (i.e. information of the cell a phone is connected to, even if no billing relevant activity takes place) is available. Should mobile internet achieve similar penetration rates in the future as do mobile phones today or if handover data becomes more widely available, these methods can be applied to the full breadth of CDR data. While some work has been put into reconstructing movement from CDR requiring less data, it mostly still interprets the data similar to GPS trajectories (Doyle et al. 2014; Schulz, Bothe, and Körner 2012; Calabrese et al. 2013). However, we think that this should only be done if the CDR are temporally dense enough. Other work on CDR has focussed on extracting derived information, such as home and work locations (Ahas et al. 2010; Becker et al. 2013; Eagle, De Montjoye, and Bettencourt 2009; Isaacman et al. 2011) or classifications of users (Furletti et al. 2013; Nilbe, Ahas, and Slim 2014). In their studies, (Schulz, Bothe, and Körner 2012; Ranjan et al. 2012) have found that using CDR as a sparsely and inaccurately sampled trajectory incurs biases, especially for users that do not have many CDR distributed evenly through the day. Possible correlations between the rate at which CDR are produced and whatever is being studied make excluding users based on their CDR rate highly undesirable. Methods that help reducing the amount of required CDR and that do not make strong assumptions on the users' behaviour are missing and the current work attempts at filling this gap.

3. Data and Preprocessing

We use both a real life dataset that shows the behaviour of the methods on messy human data as well as a simulated dataset that allows for more control over certain parameters. Section 3.1 will describe the real world dataset where as the simulation is described in Section 3.2.

3.1 *Real world dataset*

The data we use to test our methods comes from two sources. First we have information that was gathered from 140 Estonian participants during 2015 using the YouSense[1] application first presented in Linnap and Rice (2014) that was since developed further. The data comprises information on 22943 days of the users, an average of nearly 200 days per user. The collected information includes GPS positions, timestamps of sent and received text messages and calls as well as the connectivity status of the phone (i.e. what mast the phone was connected to at any given point in time).

YouSense: The GPS information comes at a sampling rate of mostly one minute, which is adequate as GPS information is only used for evaluation purposes in this

[1] http://positioner.ut.ee/dashboard/info/

study. Three reasons can lead to diversions from the usual sampling rate: The users were allowed to pause GPS recording temporarily, bad reception can prevent a clear GPS signal, and the app can pause recording if the phone does not move.

The application stores information about the status of the connection whenever that status changes (connected, flight mode, emergency calls only, no connection) along with the ID of the cell (if connected). We will refer to this information as handover data even if strictly speaking it is a bit richer due to the information beyond the simple cell ID.

The CDR are annotated with a time stamp and and the nature of the record (e.g. incoming call, outgoing text message). On average a user recorded 5.2 CDRs a day, with a positively skewed distribution with 10 % and 90 % quantiles of the users' averages of 1.9 and 10.3 respectively. For more detailed information please consult the supplementary material.

OpenCellID: In order to bring together the GPS coordinates with the information on the cell towers we used the information from a second source, namely OpenCellID[1], which are incomplete and at times inaccurate, due to their nature as volunteered geographic information (VGI) (Goodchild 2007). A summary of the different kinds of data we use can be found in Table 1.

Content	Source	Use	Description
GPS	YouSense	Ground Truth	Sampled at most once per minute.
Handover data	YouSense	"Ground Truth" at cell granularity	Connection to all the masts (even if no CDR is produced). Indication if no connection was possible
CDR	YouSense	Input for the extraction of typical days	Time and type of all CDR activities.
Cell locations	OpenCellID	Connect IDs and locations	VGI

Table 1. Information on the data used in this work

3.1.1 Preprocessing

We pre-processed the GPS data slightly to allow for easier and more reasonable comparisons. We disregarded the time between two GPS recordings that were spatially and temporally far apart (500 m and 5 min respectively) or if the temporal distance was very large (greater than two days). Next we smoothed the trajectories where they showed indoor behaviour and flagged as stop every fix which has neighbouring fixes in a contiguous time interval of at least five minutes in which there is no GPS signal outside a 100 m radius around the measurement. The next step flags hitherto unflagged points if the containing sequence of unflagged points is "short" (thresholds based on total distance travelled, total time, circle radius and number of points). The segments were then sequences of points with the same flag status.

A first informal look at the data revealed that the number of people without a clear work location and/or irregular movement behaviour was larger than expected. In the context of the current work, this is not detrimental to our findings as we would like our method to work well even for users with unconventional yet regular behaviours. More examples of users can be found in the supplementary material.

[1] www.opencellid.org

Statistic	GPS	GSM	GPS^S	$GPS^{W,S}$	CDR	CDR^W
RoG	10.2	10.4	10.9	9.9	6.2	6.8
Dist. travelled	52.6	221.3	35.6	35.6	23.1	23.1

Table 2. Average over all the user days of radii of gyration (RoG) and distance travelled in km on the three available levels of the data: GPS, Handover and CDR. W stands for numbers using time weighted points, S stands for GPS calculations on stops only.

3.1.2 Measures of information loss of CDR data

In order to justify our reservations towards treating CDR as a normal movement trajectory, we would like to present some summary statistics that were also used in (Schulz, Bothe, and Körner 2012; Ranjan et al. 2012). We have calculated both the distance travelled and the weighted as well as the unweighted radii of gyration (RoG)

$$r_g := \sqrt{\frac{\sum_{i=1}^{N} w_i \cdot (p_i - \bar{p})^2}{\sum_{i=1}^{N} w_i}} \qquad \text{with} \qquad \bar{p} := \frac{\sum_{i_1}^{N} w_i \cdot p_i}{\sum_{i=1}^{N} w_i}$$

with N the number of points and w_i the time spent in p_i for the weighted version and $w_i \equiv 1$ for the unweighted version. There is no weighting equivalent for the distances so the same results are shown for both versions in Table 2.

We compare the following: GPS trajectories (GPS), handover data seen as trajectories (GSM), the stops of the segmented GPS trajectories with equal weights (GPS^S) and with weights based on stay times ($GPS^{W,S}$) and CDR data seen as trajectories, again using equal (CDR) or time dependent (CDR^W) weights. In the latter case the weight of a CDR was chosen as the difference in time between the temporal midpoint between the CDR and its successor and the midpoint between the CDR and its predecessor. This should help eliminate the influence of bursts on RoG.

GSM clearly and unsurprisingly overestimates distance travelled, it fares rather well in estimating the RoG.

Due to the typically small amount of time spent on move segments, the RoG of GPS^S are close to the GPS radii. On the other hand the distance travelled as measured by GPS^S is below GPS, as the assumption of movement in a straight line between stops is clearly a simplification.

CDR and CDR^W are both unable to capture any of the measures due to movement without CDR which means that viewing CDR data as trajectories cannot lead to adequate results, at least in our dataset (cf. also the supplementary material). We therefore must look for alternative ways to capture movement from CDR data.

3.2 *Simulation study*

Real world data can provide a sanity check under messy real world conditions. However it can also be helpful to evaluate and compare different methods also in the tightly controlled environment of a simulation to clearly identify circumstances where a method works particularly well or badly. We therefore simulate the (regular parts) of the movement of people using the masts from a 10 km radius around Tartu. To avoid the necessity of first clustering masts we choose the masts such that they are at least a certain distance apart from each other (1 km and 500 m respectively, depending on the number of drawn masts). For the association mining approach the actual position is of no importance, as the labels do not carry any

geographic information. For the method we propose the distance between masts matters and therefore a mostly urban environment seemed appropriate, as the majority of the participants of the study comes from urban areas. Next we randomly create routines for each user, each consisting of a morning, an afternoon and an evening activity with the rest of the time being spent at a randomly drawn home location. Those routines are created directly in vectorised form, such that we know the upper limit to reach is zero deviation. We then create 200 sample days from those routines (the number reflects the real life participants), adding uncertainty about the exact beginning and ending of each of the activities to allow for uncertainty. The underlying routine for every sample day is selected among the available routines for a user with exponentially decaying probabilities in order to reflect the fact that some routines are more frequent than others. Every of those simulated locations is observed with a probability that is proportional to a linear combination of the observed hourly CDR frequencies and a constant probability for every hour.

For every combination of the following parameter choices, 20 users are simulated with 200 days each resulting in $2^4 \cdot 20 \cdot 200 = 64'000$ simulated days:

- Number of Locations: Either 5 or 15 masts are used as pool from which to generate routines. 5 is chosen as the upper bound of the very few locations that most users seem to spend most of their time according to (Bayir, Demirbas, and Eagle 2010), 15 is a number large enough to allow most routines to happen in (almost) disjunct non-home locations.
- Number of Routines: Either 2 or 4 Routines are generated. 2 To reflect a Weekday-Weekend dichotomy, and twice as much, to add more complex behaviour.
- Calling Probabilities: Either the empirical probabilities for CDRs (EP) or $0.6EP + 0.4 \cdot 1/24$ are taken as base (scaled to sum to 1 over a day). The linear combination was chosen to see whether the first and last locations are fitted better when the CDRs are more dispersed.
- Factors: The base probabilities are then multiplied by 3 or 6, resulting in expected CDR counts of 3 and 6 respectively. The choice for 6 (on average) was taken as half of what (Pappalardo et al. 2013; Becker et al. 2013; Isaacman et al. 2011) used or had, as we specifically want to use methods that work on moderate counts of daily CDR. We then halve that again, to see how far down we can go.

4. Methods

There is a consensus that daily human mobility patterns show a high regularity (Lu et al. 2013; Song et al. 2010; Schneider et al. 2013). A reasonable assumption therefore would seem that this regularity, once learned, should be conducive to the quality of reconstructing the whereabouts of mobile phone users.

4.1 *Representation of the data*

CDR activity is "bursty" (Barabasi 2005; Gonzalez, Hidalgo, and Barabási 2008; Song et al. 2010), with a considerable number of CDRs happening in close temporal proximity of others. This can result in an overrepresentation of certain cell-towers in the data. A possible representation to solve this issue can e.g. be found in (Furletti et al. 2012): The day is partitioned into equally sized intervals and the CDR is recorded in the interval in which it happened. As every time slot has the capacity

for only one piece of information, only one CDR can be considered; we choose the one closest to the center of the time slot. This way, most of the members of the bursts are binned together. Should the burst fall right on the (arbitrary) border of the time slots, both adjacent slots will only contain the same information if there is no other CDR in one of the slots that is closer to the respective midpoint of the slots. The resulting vector is often not complete, as e.g. with two CDRs in a day, at most 2 time slots may be filled. The finer the temporal partition, the sparser the resulting data vectors will be populated. There did not seem to be any natural best choice for the temporal granularity so we have done our analysis with partitions of the day into 24, 12 and 6 time slots. Finer partitions seemed to be too granular for the CDR counts we observed whereas coarser partitions would correspond to blocks larger than 4 hours which already seem at the border of what seems sensible.

The spatial location is measured on the granularity of the cell regions. Due to the nature of the connections it makes sense to cluster cells that were frequently used and that are close together into a location that is meaningful to the user. The approach chosen by Do and Gatica-Perez (2012) uses equally spaced grids for the analysis and thus does not take into account that the density of the cell towers varies by at least an order of magnitude between urban and rural settings. On the other hand, Csáji et al. (2013) uses twice the maximum distance of a cell to its Voronoi-neighbours to cluster the points. While we have no information about how the factor of two was calculated, if we look at our data, this factor seems to be dependent on the distances to the Voronoi neighbours, as illustrated in Figure 1. There we plot the average distance of a cell location to its Voronoi-neighbours against the multiple of that distance required to ensure at least a given quantile of the GPS fixes that were recorded while connected to a mast lie within a circle of that radius around the corresponding cell centre's location. The results for the multiples of the maximum (in stead of average) distances to the Voronoi-neighbours are very similar, so we take the average instead, as it is slightly more resilient against outliers than the maximum. The multiples required are of course noisy due to the limited number of users, as cells that "see" few users might have their estimation dominated by the distance to the frequent location of a single user. However, there seems to be a clearly discernible linearity in the trend, as the lines drawn are in fact smoothing splines and could bend if the data suggested non-linearity. The fact that the multiplier should depend on the distance to the Voronoi-neighbour seems natural: While a cell in the inner city may easily serve a phone three cells away due to the high density of the (Voronoi) cells, a rural cell of 10 km radius may not be able to do so. When measuring distances from cell centres, we therefore use scaled versions of those distances, i.e. we divide them by the expected radius of the circle containing 75 % of GPS points. For ranges between 50 % and about 80 %, different choices for the threshold scale the adjusted distances approximately linearly, which can be fully compensated by the clustering that follows, so within this range any value can be chosen. We retained 75 %.

We then calculate the rescaled distance matrices between the used cell centres for each user and use them as input to DBSCAN (Ester et al. 1996) to find clusters. The IDs of cells from those clusters are then changed to the corresponding cluster ID and the location of the cluster are set to the mean of the locations of the contributing cells' centres. Apart from identifying potentially semantically meaningful places of a person, this has the advantage of reducing the number of recorded "cells", facilitating the recognition of patterns.

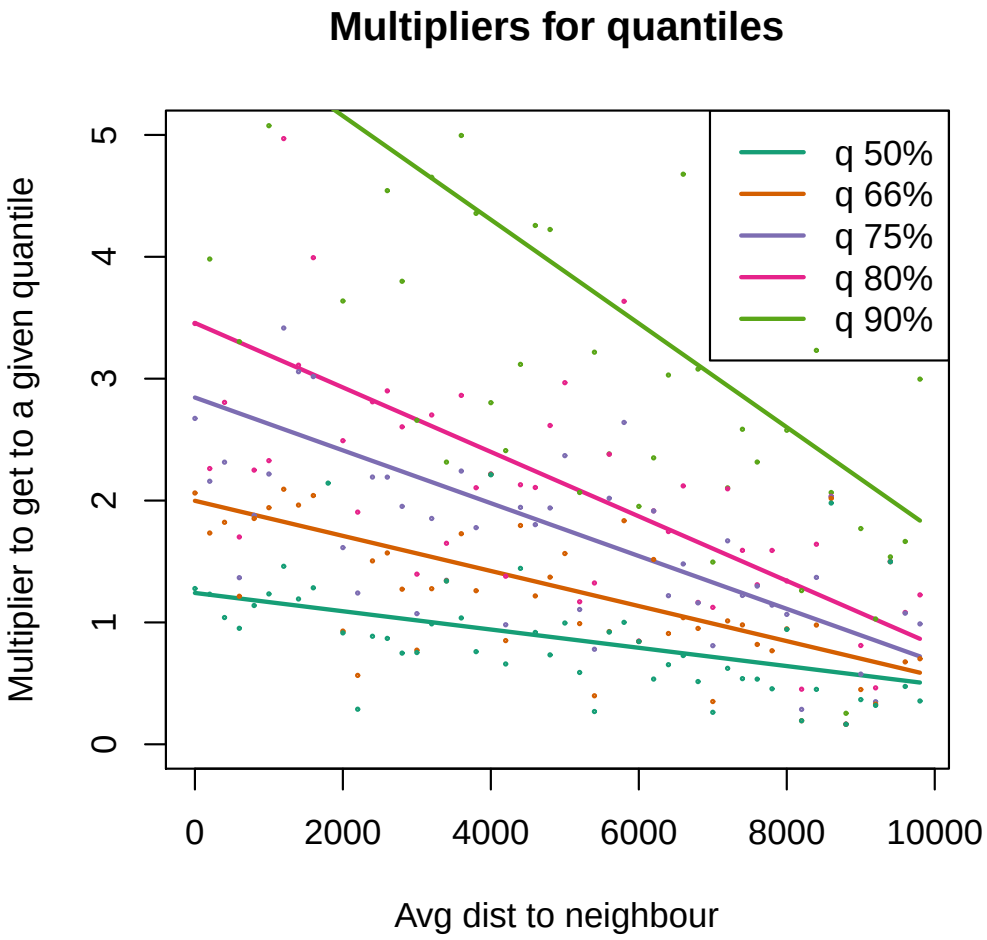


Figure 1. Multiples of the average distance to the Voronoi-neighbours of a cell that describe the necessary radius of a circle around a cell centre required to capture at least a certain quantile of the GPS locations that were recorded while connected to the mast.

4.2 *Reconstructing trajectories*

We now present the different methods for reconstructing trajectories from CDR data.

4.2.1 Method 1: Association Mining

The first method is the mining of association rules using the *apriori* algorithm (Hahsler, Grün, and Hornik 2005) with a combination of cell ID and time slot as input. We set a low support threshold (2 items) to get broad range of potential rules. Given a sample day with the recorded CDRs in their respective time slots as left hand side we look for the rule with the highest lift for every missing time slot and fill the gaps in this way. Time slots that have no rules given the observations are filled with the closest available information after the rules have been applied. This typically is the case for the very early and very late time slots, that are then simply filled with the first/last predicted location. The advantage of this approach is that it is relatively stable and can deal with different amounts of CDR: the more data it is fed, the more nuanced the rules can become. On the other hand it does not embed any notion of temporal proximity, as the items are just (uninterpreted)

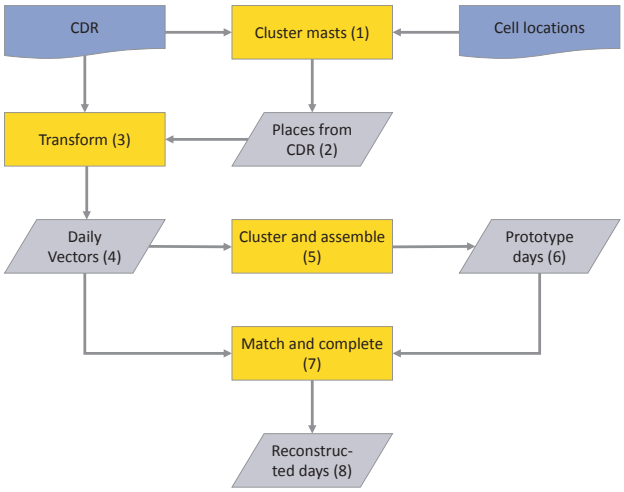


Figure 2. Workflow of the DAMOCLES approach.

labels and the rules it finds are again on a label level and the method does not produce representative days.

4.2.2 Method 2: DAMOCLES approach

As an alternative we propose DAMOCLES, the *DAily MObility CLustering for Extracting Space usage*. The overall idea is to use the temporal regularity of human movement behaviour to identify typical movement behaviour. In essence this attempts to compensate low CDR counts with aggregation over time. We identify days that are similar and use them to create prototype days, which then can be used to complete sparsely populated daily CDR vectors. As the first method, DAMOCLES is an approach that searches recurring patterns in the existing data. However, in contrast to association mining where all the available items are independent from each other and do not themselves carry spatial or temporal information, it explicitly captures the spatial and temporal structure of the data. There are three main parts to the algorithm, whose schematic representation can be found in Figure 2:

(1) A dissimilarity measure,
(2) a clustering algorithm and
(3) a reconstruction based on the identified clusters.

Specifically, given the set of cells $\mathfrak{C}$ and the number of time slots $n_t \in \mathbb{N}$ that we divide a day into, we define the extended set of cells $\mathfrak{C}_e := \mathfrak{C} \cup \{NA\}$, the set of days $\mathfrak{D} := \mathfrak{C}^{nt}$ and the extended set of days $\mathfrak{D}_\mathfrak{e} := {\mathfrak{C}_\mathfrak{e}}^{n_t}$. The extended versions are needed, as the observations may contain missing values. We observe $n_d \in \mathbb{N}$ days $D^o \subset {D_e}^{n_d}$ (4). We then use a dissimilarity function $d : D_e \times D_e \to \mathbb{R}_0^+$, a clustering method $c : {D_e}^{n_d} \to \mathbb{N}^{n_d}$ and a cluster assembly method $a : {\mathfrak{D}_\mathfrak{e}}^n \to \mathfrak{D}$ for some $n \in \mathbb{N}$ (5). Lastly we need a reconstruction $r : {D_e}^{n_d} \times \mathbb{N}^{n_d} \to D^{n_d}; \quad D^o \times c(D^o) \mapsto D^r$ where D^r denotes the reconstructed days (6).

Implementation

For the clustering (5), we opted for DBSCAN, as it allows for different numbers of identified clusters per user and can accommodate users with different number of recorded days. Having a method that allows for different numbers of clusters is required as some users might simply have a weekday and a weekend routine

whereas others might show more diverse regular days. DBSCAN needs as input a dissimilarity matrix with pairwise dissimilarities between the entities that need to be clustered. This dissimilarity matrix is calculated using d, which needs to fulfil positive semi-definiteness and symmetry in order for DBSCAN to yield sensible results. Note that it does not need to be a metric, as DBSCAN can cope with d fulfilling neither the triangle inequality nor the identity of indiscernibles. Special care should be given to how the dissimilarity treats the missing values: One has to avoid DBSCAN connecting everything through (almost) empty observations. We set:

$$d(day^{(1)}, day^{(2)}) \coloneqq$$
$$\sigma\Bigg(\sum_{i=1}^{n_t} \min_{t_{i,1},t_{i,2}\in\{-1,0,1\}} \Big\{d^c\left(day_i^{(1)}, day_{i+t_{i,1}}^{(2)}\right) + \frac{|t_{i,1}|}{2} + d^c\left(day_{i+t_{i,2}}^{(1)}, day_i^{(2)}\right) + \frac{|t_{i,2}|}{2}\Big\}\Bigg)$$

where $\sigma(x) \coloneqq 1/(1+e^{-x})$. The distance measure for cells d^c uses a combination of the Euclidian distance d^c and the adjusted distance that we used in the clustering. Negative values bring the distance d between the days closer to zero, whereas positive values bring it closer to one.

In d^c we want negative values if the cells are the same or at least very close. If there is no overlap (low probability of the person being at the same location but being connected to two different cells), we want to penalise according to the distance: As larger differences in a specific time slot make it less likely that the difference is due to a slight deviation from a normal pattern, we want to penalise larger distances stronger than small distances. All of the above resulted in the following definition for d^c:

$$d^c(c_1, c_2) \coloneqq \begin{cases} -1 & \text{mutual overlap} \\ -0.5 & \text{one sided overlap} \\ NA & \text{one of the cells is NA} \\ log_{100} d^e(c_1, c_2) & \text{otherwise} \end{cases}$$

The minimum in d treats NA as plus infinity. If one of the two parts in the minimum cannot be brought to a real value (i.e. all timeslots in a 1-neighbourhood are missing values), we set the term to zero. Overlap happens if the second cell in question has a Euclidian distance to the first cell that is less than what could be expected based on what we learned from Figure 1. This formulation of the distance is very much related to localised dynamic time warping (Berndt and Clifford 1994) in that we are looking for a least-cost path through pairs of cells. The difference however lies in the way this formulation lets us treat missing values. If we give a reasonable (i.e. close to 0) cost to a connection to a missing value directly in d^c then timeslots with far away cells are avoided in favour of empty cells in the matching process, making the days seem more similar than they should. We therefore allow connections to missing time slots only when there is no available observation in the whole 1-neighbourhood. The proposed distance measure for days is both positive and symmetric, which are the requirements for DBSCAN.

One generally would like to have a small epsilon environment (only cluster days

that have matching cells in many time slots) but certain clusters simply are not discernible at too small thresholds. Choosing the threshold too large on the other hand creates the risk of not distinguishing between different clusters or clustering days that do not at all represent similar days. To overcome this issue, we iteratively apply DBSCAN with a sequence of increasing ϵ and remove days belonging to identified clusters from the set of days to consider (Cf. Supplementary material for a more detailed explanation).

The reconstruction of days (7) then is fairly simple: Given the observations of a day we look for the cluster that is the closest (again using d) to the observation. If there are multiple candidates, we take the one with the lowest cluster number, corresponding to the cluster with the smallest epsilon environment. From that cluster we take the mode of cells at every time slot, removing those time slots where the mode appears only once (typically early in the morning or late at night). We then use that information to fill in the missing values of the observation. Any time slots that are still missing are then filled by the closest non-missing value.

4.2.3 Evaluation

We compared the two proposed methods with two simple benchmarks: The first one (denoted "mode by slot") assigns the most frequently seen cell by time slot to the time slots with no observation (i.e. one cluster over all observed days). The second one (denoted mode by time and Weekday/-end) assigns the mode of the cells observed by time slot and an indicator function for Weekends (Saturdays and Sunday) to missing observations (i.e. clusters follow days of the week), was implicitly or explicitly assumed in (Jiang, Ferreira, and Gonzalez 2012; Kung et al. 2014; Ahas et al. 2015; Ranjan et al. 2012).

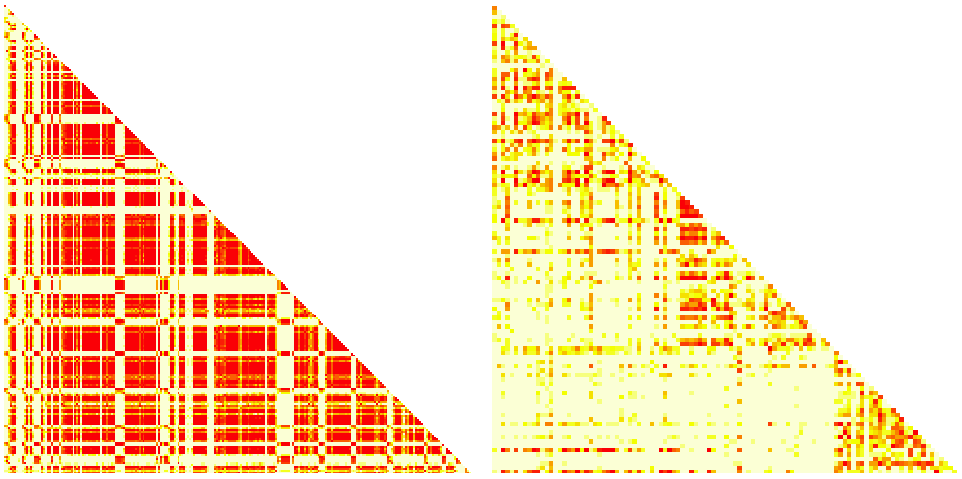
For the evaluation and comparison of the different methods to identify idiosyncratic daily behaviours we calculate and compare the distances between predictions and the actually recorded positions. The predicted location remains constant for every predicted time slot and thus is both spatially (cell size) as well as temporally imprecise. As the GPS measurements sometimes come at irregular intervals all measurements are weighted by the durations of the intervals during which the GPS-position was not updated.

To put the obtained results into perspective we also calculate the distances obtained by using the handover data in temporal segments that reflect the actual connection (i.e. not matched to time slots). This sets a natural upper limit to the accuracy of the predictions. As we use clustered cells, it is possible that the centroid of the cluster is closer to e.g. the home of a user than any of the individual masts, so it can happen that the prediction has a lower average distance than the cell tower "ground truth".

5. Results

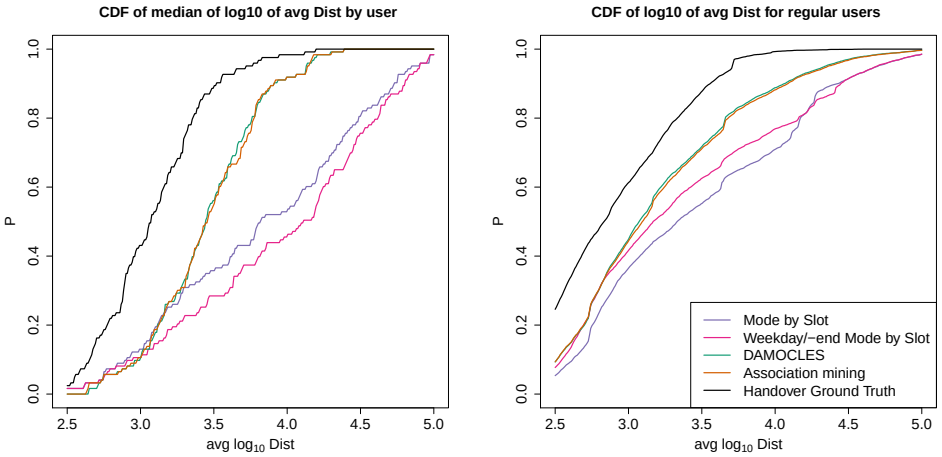
5.1 *Experimental data*

In order for the clustering to work, the distance matrix between days of a user should show a partition of the days into groups whose members are close to each other and far away from members of other groups. Examples for our users can be seen in Figure 3. In the distance matrix on the left the users' days partition nicely into two very clear groups. On the right hand side one can see a user with three regimes that follow one another. This indicates that the chosen distance measure is capable of distinguishing patterns belonging to daily routines that happen



(a) User 187 (b) User 116

Figure 3. Distance matrices between the days for two users, with the distance between days i and j in row i and column j. The brighter the colour the greater the distance between two days. On the left hand side, two different regimes are very clearly visible. On the right hand side, there are three regimes that follow one another and are separate from each other.



(a) All users (b) Users with regular behaviour

Figure 4. For all users: Cumulative distribution function (CDF) of the logarithm of the average distance between the prediction and the ground truth by user. For users with a regular behaviour: CDF of the average daily distances.

regularly.

Figure 4 shows for all users the CDF of the median of the daily average Euclidian differences for the different methods by user. This way every user gets equal weight, irrespective of the number of days she was under study. Clearly the two methods that we propose are better at reconstructing the actual movement of the users than the benchmark solutions, indicating that the patterns captured by them are more helpful for estimating the users' whereabouts. Note that even if we take the handover ground truth, there are days where the average distance is considerable, hinting at an irreducible uncertainty that comes with the data and can come from time spent in regions where the cells were large or from incorrectly geo-referenced cells in OpenCellID. The general appearance of the image is the same for all three tested partitions of the day, so we only show the one corresponding to 12 partitions.

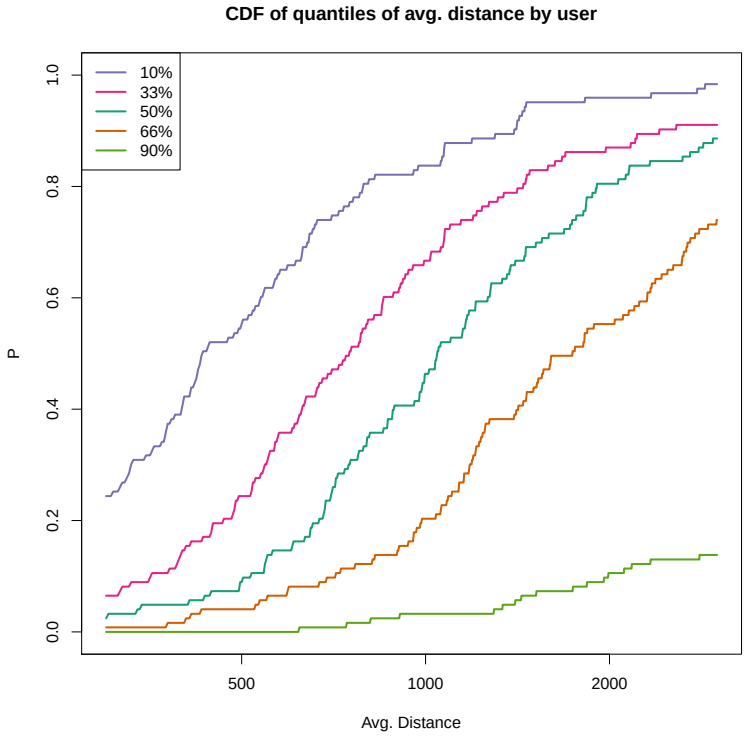


Figure 5. Cumulative distribution function of different quantiles of average daily distances by user for the DAMOCLES method.

Of course the median cannot convey the behaviour of the prediction in all detail. To shed some additional light we also show the behaviour for quantiles other than the median in Figure 5, but limited to the DAMOCLES method to avoid overcrowding the image. The other methods show fans of similar width around their respective median curves (not shown). For almost all the users there seem to be at least 10 % of days that are very poorly predicted irrespective of method and number of time slots (not shown). We feel that the exact number is of minor importance, as the users under study are not representative of any general population and rather small in number. These days can correspond to larger trips without CDR that may start or end at home and thus can be falsely attributed to a prototype day leading to grossly wrong predictions. Other causes can be a high proportion of movement throughout a day, such that the restriction of the prediction to the time slots of fixed width prohibits an adequate representation of the movement or they can be the result of locations that were never recorded by the CDR. An example of the latter can be seen in Figure 6 where there is a clearly discernible frequent location that is visited after what can be presumed to be work, but where there is no CDR that would allow us to capture this behaviour in the first quarter of 2015 (and only very few in the rest of the year).

We have included some images depicting examples of reconstructions in geographical space in the supporting material.

Our sample size is limited, so subdividing the population into sub-populations (such as frequent and infrequent callers) leads to results that strongly vary with the individuals, so we do not make many statements about subpopulations. One that we would like to make however is one about users that show a particularly regular user behaviour. In our sample, all users with a very regular behaviour worked during the day at more or less fixed times from Monday to Friday. For those users

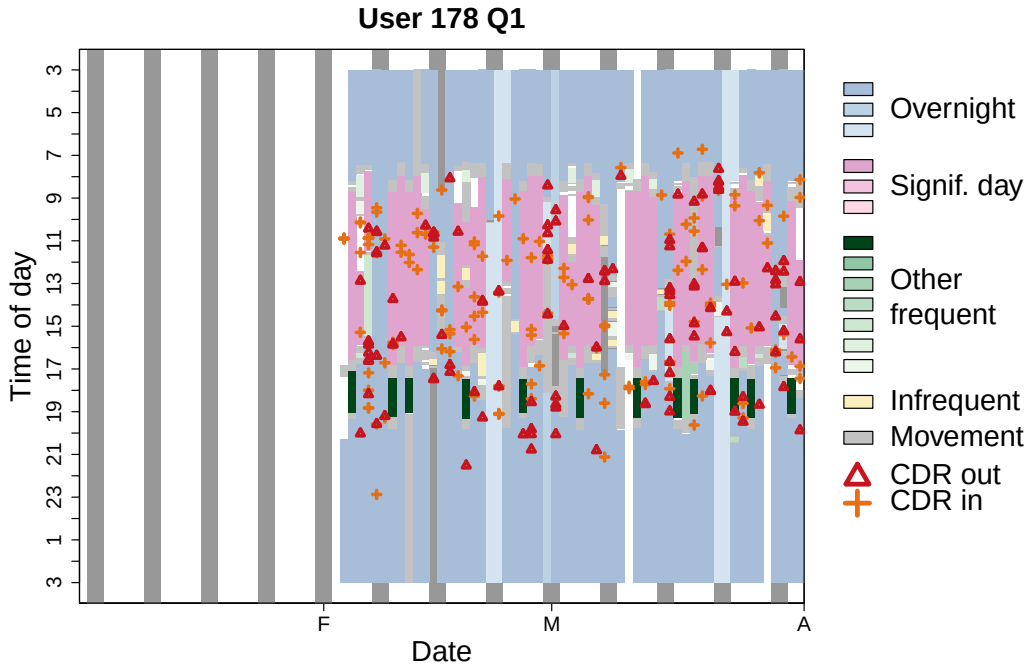


Figure 6. Example of a user with a frequent location (in dark, saturated green, between roughly 17:00 and 19:00) where there is never any CDR activity. A description of how to extract the full set of information from this plot can be found in the supporting material.

we would expect the weekday / weekend benchmark to perform rather well, and indeed it does, as can be seen in Figure 4. Note that both methods that do not make a priori assumptions are at least as good as the benchmarks even if the users happen to actually work on exactly the days that the benchmarks assume.

Neither the association mining approach nor DAMOCLES is clearly stronger for all users on the data collected by YouSense. You can find some visual examples of reconstructions in the supporting material.

5.2 *Simulation*

The results of the simulation study can be found in Figure 7. interestingly DAMOCLES seems to work better than the association mining approach on days where there are just very few CDR. Most strikingly this is the case for label correctness, where the median lies at nearly 50 %, whereas association mining has a median of just over 30 %. While less pronounced, similar observations can be made of the reconstruction error, measured by average distance between simulation and reconstruction, where DAMOCLES is less error prone on days with low CDR counts. In both cases the methods start to look similar as soon as 5-6 CDRs are recorded on a day. This also happens to be the approximate value of the threshold needed to reach the saturation point, at which the error reaches the (irreducible) error incurred by the randomly fluctuating starting times of the activities.

The results of Figure 7 are qualitatively similar if we subset the total simulation population into the classes identified by the choices for the parameters. However, we would like to compare the overall picture with the one obtained from the subpopulation that had 3 CDRs a day on average that you can see as Figure 8. We can see that both approaches still can yield reasonable results for users with CDRs in as few as 3 time slots a day on average.

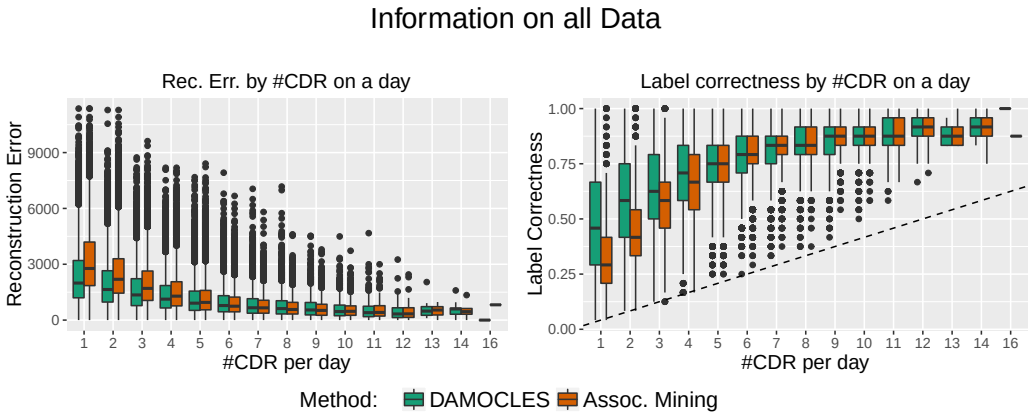


Figure 7. Boxplots with the results for the simulation study. Left: the reconstruction error as daily averages of distances between simulated and reconstructed locations. Right: Daily averages of correctly attributed mast IDs.

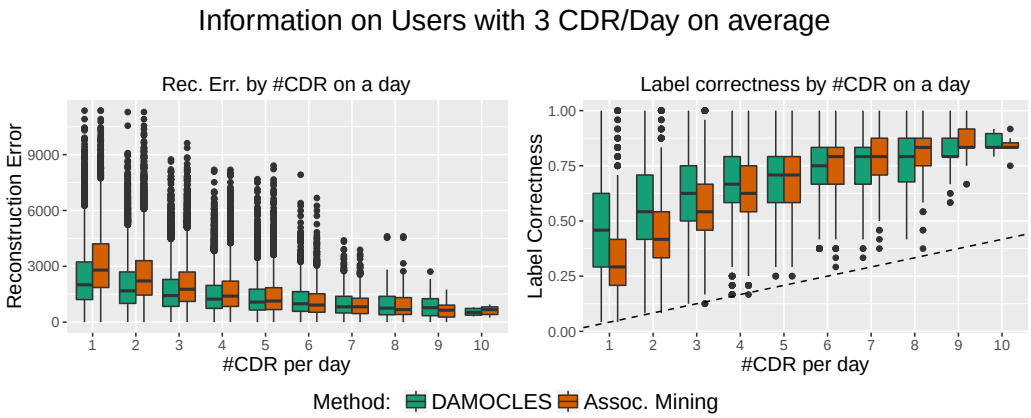


Figure 8. Boxplots with the results for the simulation study for users with 3 CDRs a day on average. Left: the reconstruction error as daily averages of distances between simulated and reconstructed locations. Right: Daily averages of correctly attributed mast IDs.

6. Discussion and Conclusion

In the current work we could show that association mining and DAMOCLES can both be used to reconstruct daily whereabouts of users, given their CDRs for an extended period of time. Both methods can capture different habits of movement in ways that do not require a priori assumptions on working days (Jiang, Ferreira, and Gonzalez 2012) or working hours (Ranjan et al. 2012).

Association mining is computationally fast and yields stable results, as shown in the study on the data collected via the YouSense application. However this method is not able to capture the spatio-temporal information underlying the data, and specifically does not distinguish between a small or large spatial error. Also, all the rules learned concern only individual time slots and the big picture of what a typical day as a whole looks like is missing.

DAMOCLES on the other hand is able to find examples of typical days in a way that considers the spatio-temporal characteristics of the underlying data in its decisions. In addition, whole days are considered, which allows for a more interpretable result as well as a superior reconstruction performance on days with only few observed CDRs, as the simulation study demonstrates.

The absence of *a priori* assumptions means that both tested methods yield their

results irrespective of working hours or the days of the week that the movement habits follow. This is a clear advantage, as demonstrated by the worse performance of the benchmarks representing those assumptions in the YouSense study. This can be seen as an indicator that these methods can be used for studying large fractions of a population, where systematic errors on night or weekend workers may bias the findings.

Apart from its benefits, DAMOCLES also has its limitations. Due to the clustering, it can only work if there are enough days in which there are enough (and dispersed enough) CDRs that allow the distance function to get low enough for clustering. This means that the method does not work for users with constantly very low numbers of CDR. However, as the "high enough" number only are needed to identify the clusters, the average number of CDR per day can be much lower than for methods that directly reconstruct movement from CDR data (Widhalm et al. 2015; Schulz, Bothe, and Körner 2012). Specifically, the simulation study shows that we get reasonable results for users with as few as three CDR per day on average. Another limitation that is inherent in CDR data is that it can only capture locations where CDRs occur and hence any unreported locations will be missed.

There are several ways in which DAMOCLES could be extended. As it is presented here, the temporal regularity of the typical days is not used in order to reduce the assumptions made to a minimum. If one is willing to assume that there is some regularity in the temporal sequence of daily regimes one could easily extend both the clustering and the matching parts of the algorithm to include information on e.g. CDRs from preceding and succeeding days or the day of the week. A second extension that could benefit both DAMOCLES and the association mining approach concerns the first and last location on a day. Some users hardly ever have CDRs in the GPS stop segment that covers midnight and therefore both approaches at times fail to detect the first and last stop segments of a day. One way of dealing with this issue could be to include any of the methods from the literature to find sleeping locations (e.g. (Ahas et al. 2010)) and select the first and last locations based on the estimated probabilities of the identified locations.

Lastly, one could develop an integrated approach that combines methods for different amounts of information to reconstruct every day as well as possible. For time spans with high CDR counts one could go for a method as fine grained as (Widhalm et al. 2015), whereas for intervals with fewer observed CDRs, one could use e.g. DAMOCLES. To extract the intervals on which to use the more refined method, a sensitivity study on that method that detects when it breaks down would be necessary.

We are convinced that models with few a priori assumptions about human mobility are needed when large parts of the population are analysed. Especially minority populations that do not conform to standard assumptions about everyday habits may otherwise be misrepresented. We have contributed one such method and look forward to further research in this direction.

References

Ahas, R., A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru, and M. Zook. 2015. "Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn." *International Journal of Geographical Information Science* 8816 (July): 1–23.

Ahas, Rein, Siiri Silm, Olle Järv, Erki Saluveer, and Margus Tiru. 2010. "Using Mobile

Positioning Data to Model Locations Meaningful to Users of Mobile Phones." *Journal of Urban Technology* 17 (1): 3–27.

Barabasi, Albert-Laszlo. 2005. "The origin of bursts and heavy tails in human dynamics." *Nature* 435 (7039): 207–211.

Bayir, Murat Ali, Murat Demirbas, and Nathan Eagle. 2010. "Mobility profiler: A framework for discovering mobility profiles of cell phone users." *Pervasive and Mobile Computing* 6 (4): 435–454.

Becker, R.a., R. Cáceres, K. Hanson, S. Isaacman, J.M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. 2013. "Human mobility characterization from cellular network data." *Communications of the ACM* 56 (1): 74.

Becker, Richard A., Ramón Cáceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2011. "A tale of one city: Using cellular network data for urban planning." *IEEE Pervasive Computing* 10 (4): 18—-26.

Berndt, Donald, and James Clifford. 1994. "Using dynamic time warping to find patterns in time series." *Workshop on Knowledge Knowledge Discovery in Databases* 398: 359–370.

Blondel, Vincent D, Adeline Decuyper, and Gautier Krings. 2015. "A survey of results on mobile phone datasets analysis." *Arxiv preprint* arXiv:1502.03406v1.

Calabrese, Francesco, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira, and Carlo Ratti. 2013. "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example." *Transportation Research Part C: Emerging Technologies* 26: 301–313.

Csáji, Balázs Cs, Arnaud Browet, V. a. Traag, Jean Charles Delvenne, Etienne Huens, Paul Van Dooren, Zbigniew Smoreda, and Vincent D. Blondel. 2013. "Exploring the mobility of mobile phone users." *Physica A: Statistical Mechanics and its Applications* 392 (6): 1459–1473.

Do, Trinh Minh Tri, and Daniel Gatica-Perez. 2012. "Contextual conditional models for smartphone-based human mobility prediction." *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)* 163.

Doyle, John, Peter Hung, Ronan Farrell, and Seán McLoone. 2014. "Population Mobility Dynamics Estimated from Mobile Telephony Data." *Journal of Urban Technology* 21 (2): 109–132.

Eagle, Nathan, Yves Alexandre De Montjoye, and L. M A Bettencourt. 2009. "Community computing: Comparisons between rural and urban societies using mobile phone data." *Proceedings - 12th IEEE International Conference on Computational Science and Engineering, CSE 2009* 4: 144–150.

Ester, Martin, Hans P Kriegel, Jorg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *Second International Conference on Knowledge Discovery and Data Mining* 226–231.

Furletti, Barbara, Lorenzo Gabrielli, Chiara Renso, and Salvatore Rinzivillo. 2013. "Analysis of GSM calls data for understanding user mobility behavior." *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013* 550–555.

Furletti, B., L. Gabrielli, S. Rinzivillo, and C. Renso. 2012. "Identifying users profiles from mobile calls habits." *Proceedings of the ACM SIGKDD International Workshop on Urban Computing . ACM.* 17–24.

Gonzalez, Marta C, Cesar A Hidalgo, and Albert-László Barabási. 2008. "Understanding individual human mobility patterns." *Nature* 453: 779–782.

Goodchild, Michael F. 2007. "Citizens as sensors: The world of volunteered geography." *GeoJournal* 69 (4): 211–221.

Hahsler, Michael, Bettina Grün, and Kurt Hornik. 2005. "Association Rules and Frequent Item Sets." *Journal of Statistical Software* 14 (15).

Houston, Douglas, Thuy T. Luong, and Marlon G. Boarnet. 2014. "Tracking daily travel; Assessing discrepancies between GPS-derived and self-reported travel patterns." *Transportation Research Part C: Emerging Technologies* 48: 97–108.

Isaacman, Sibren, Richard Becker, Ramon Caceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. 2011. "Identifying important places in people's lives from cellular network data." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*

6696 LNCS (June): 133–151.

Janzen, Maxim, Maarten Vanhoof, Kay W Axhausen, and Zbigniew Smoreda. 2016. "Estimating Long-Distance Travel Demand with Mo-bile Phone Billing Data." In *16th Swiss Transport Research Conference (STRC 2016),* Swiss Transport Research Conference (STRC).

Jiang, Shan, Joseph Ferreira, and Marta C. Gonzalez. 2012. "Clustering daily patterns of human activities in the city." *Data Mining and Knowledge Discovery* 25 (3): 478–510.

Krumm, John, Dany Rouhana, and Ming Wei Chang. 2015. "Placer++: Semantic place labels beyond the visit." *2015 IEEE International Conference on Pervasive Computing and Communications, PerCom 2015* 11–19.

Kung, Kevin S., Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. 2014. "Exploring universal patterns in human home-work commuting from mobile phone data." *PLoS ONE* 9 (6).

Linnap, Mattias, and Andrew Rice. 2014. "Managed Participatory Sensing with YouSense." *Journal of Urban Technology* 21 (2): 9–26.

Louail, Thomas, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. 2014. "From mobile phone data to the spatial structure of cities.." *Scientific reports* 4: 5276.

Lu, Xin, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. 2013. "Approaching the limit of predictability in human mobility.." *Scientific reports* 3: 2923.

Nilbe, Kati, Rein Ahas, and Siiri Slim. 2014. "Evaluating the Travel Distances of Events Visitors and Regular Visitors Using Mobile Positioning Data: The Case of Estonia." *Journal of Urban Technology* 21 (2): 91–107.

Pappalardo, Luca, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, and Fosca Giannotti. 2013. "Comparing general mobility and mobility by car." In *11th Brazilian Congress on Computational Intelligence (BRICS-CCI CBIC),* 665–668.

Ranjan, G., H. Zang, Z.-L. Zhang, and J. Bolot. 2012. "Are call detail records biased for sampling human mobility?." *ACM SIGMOBILE Mobile Computing and Communications Review* 16 (3): 33.

Rinzivillo, Salvatore, Lorenzo Gabrielli, Mirco Nanni, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. 2014. "The Purpose of Motion : Learning Activities from Individual Mobility Networks." *International Conference on Data Science and Advanced Analytics (DSAA14)* .

Schneider, Christian M., Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C. González. 2013. "Unravelling daily human mobility motifs.." *Journal of the Royal Society, Interface / the Royal Society* 10 (84): 20130246.

Schulz, Daniel, Sebastian Bothe, and Christine Körner. 2012. "Human mobility from GSM data-a valid alternative to GPS?." In *Mobile data challenge 2012 workshop, June,* 18–19.

Shen, Li, and Peter R Stopher. 2014. "Review of GPS Travel Survey and GPS Data-Processing Methods." *Transport Reviews* 34 (3): 316–334.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. "Limits of predictability in human mobility.." *Science (New York, N.Y.)* 327 (5968): 1018–1021.

Steenbruggen, John, Emmanouil Tranos, and Peter Nijkamp. 2015. "Data from mobile phone operators: A tool for smarter cities?." *Telecommunications Policy* 39 (3-4): 335–346.

Tanahashi, Y., J.R. Rowland, S. North, and K.-L. Ma. 2012. "Inferring human mobility patterns from anonymized mobile communication usage." *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia - MoMM '12* 151.

Trasarti, Roberto, Ana-Maria Olteanu-Raimond, Mirco Nanni, Thomas Couronné, Barbara Furletti, Fosca Giannotti, Zbigniew Smoreda, and Cezary Ziemlicki. 2015. "Discovering urban and country dynamics from mobile phone data with spatial correlation patterns." *Telecommunications Policy* 39 (3-4): 347–362.

Vij, Akshay, and K. Shankari. 2015. "When is big data big enough? Implications of using GPS-based surveys for travel demand analysis." *Transportation Research Part C: Emerging Technologies* 56: 446–462.

Widhalm, Peter, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C. González. 2015. "Discovering urban activity patterns in cell phone data." *Transportation* 42 (4): 597–623.

Ye, Yang, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. 2009. "Mining individual life pattern based on location history." *Proceedings - IEEE International Conference on Mobile Data Management* 1–10.

Zhao, Ziliang, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen, and Ling Yin. 2016. "Understanding the bias of call detail records in human mobility research." *International Journal of Geographical Information Science* 8816 (January): 1–25.

To appear in the *Journal of Location Based Services*
Vol. 00, No. 00, Month 20XX, 1–10

Supplementary Material

(*Received 00 Month 20XX; final version received 00 Month 20XX; accepted 00 Month 20XX*)

The supplementary material compiled in this section consists mostly of additional images that support our conclusions in the paper or help getting a deeper understanding of our data.

1. Data visualisation

The visualisation in Figure S1 is an overlay of the CDRs over the stops identified through GPS. For this the stop segments from the segmentation were clustered (DBSCAN with $\epsilon = 30\ m$ and a minimal number of 4 points) and then classified according to temporal characteristics of appearance. Places of overnight stays are shown in hues of blue, places visited for more than three hours a day on average are shown in reds and other frequent places are shown in green tones. For each of the color scales, the most visited n ($n = 2$ for blue and red locations and $n = 6$ for green locations) places have their own colour, such that e.g. all dark red rectangles correspond to the same point. Further locations of the same type share the same colour, so that light blue locations need not be the same. Infrequently visited locations are shown in yellow and movement segments are shown in gray. The CDR are grouped into active (outgoing calls and sms, represented as red triangles) and passive CDR (incoming calls and sms, represented as orange crosses). The light gray bars in the background represent the weekends. Note that this very crude colouring scheme has no impact on the numerical evaluation of our methods and simply serves to make the plots more easily readable.

2. Statistics on the data

A very important question is the amount of CDR that we can observe for a typical day of a user. While point measures are inadequate to describe the whole behaviour, averages by user shown in Figure S2 can give an idea of the order of magnitude. As we see, most users have something between three and four CDRs per day on average, with some users being significantly more active at daily CDR counts of above ten.

If three or four CDRs in a day are evenly distributed through the day this can already provide a good idea of where the person was throughout the day. However, as the peak at very low values in Figure S3 shows, CDRs tend to happen in bursts. For a fixed total number of CDRs, the concentration of information on location on short temporal intervals makes localisation more difficult the rest of the time.

In the Figure S4 one can see the number of user days on which we have information and the probability of a user pausing the GPS recording. One can see that

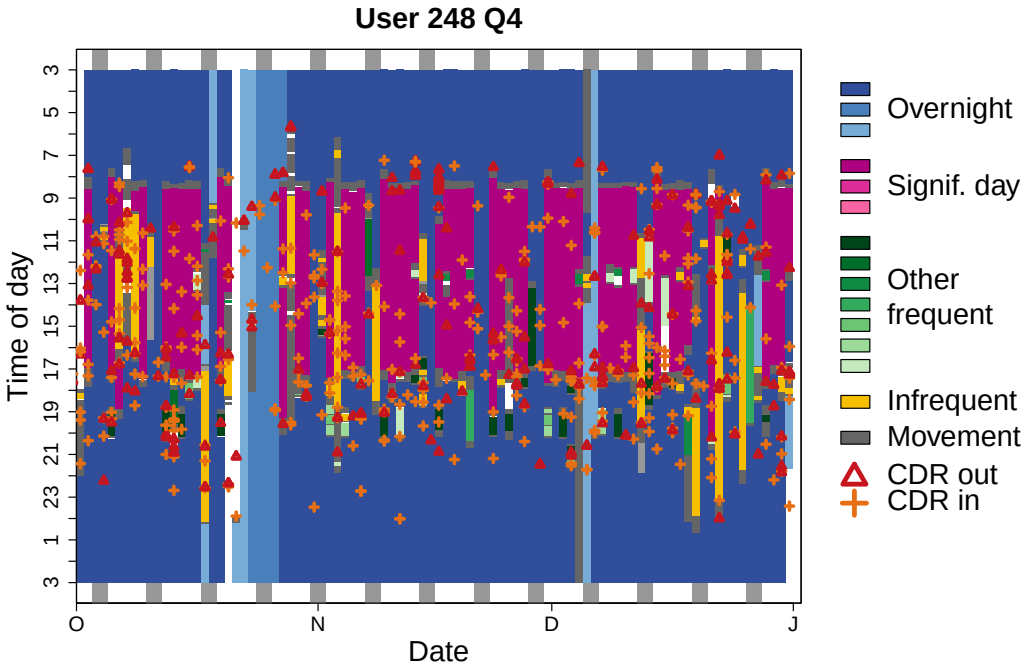


Figure S1.: Visualisation of CDR data (foreground) vis-à-vis the (interpreted) GPS data in the background. For readability's sake the plots are drawn for every user and every quarter of the year separately. This example is the final quarter for user 248.

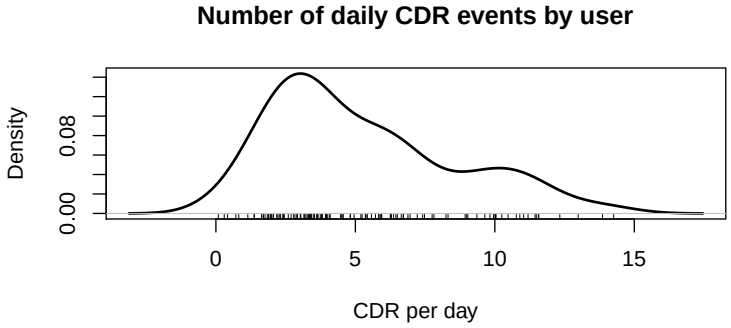


Figure S2.: Density of the average number of daily CDR per user.

for most users we have either about two hundred or above three hundred days of observation, with a smaller group of users with just about one hundred days. Contrary to our expectations, the probability of a user asking for a pause in GPS recording did not decrease in time after the first day of observation. Only at the very end of the observation period, where the number of observed users is low and thus the variance is inherently higher can we see a change in the probability.

One last thing we would like to present in this section is the density of the masts. As has been stated many times (e.g. Steenbruggen et al. (2015); Rinzivillo et al. (2012); Kung et al. (2014)), the masts are more dense in cities than in rural areas. The OpenCellID data confirms this expectation, as can be seen in Figure S5 where we plot the densities of the average distances to the Voronoi neighbours for every

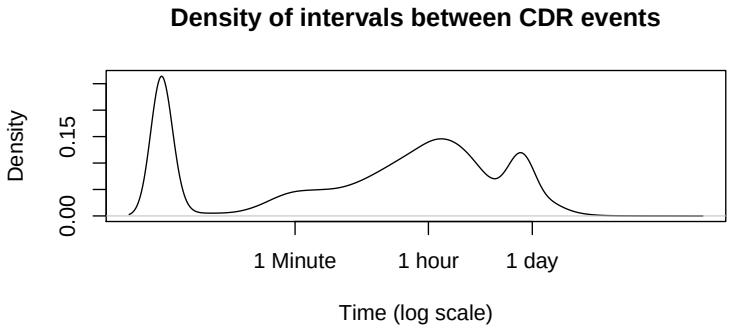


Figure S3.: Density of time between consecutive CDRs.

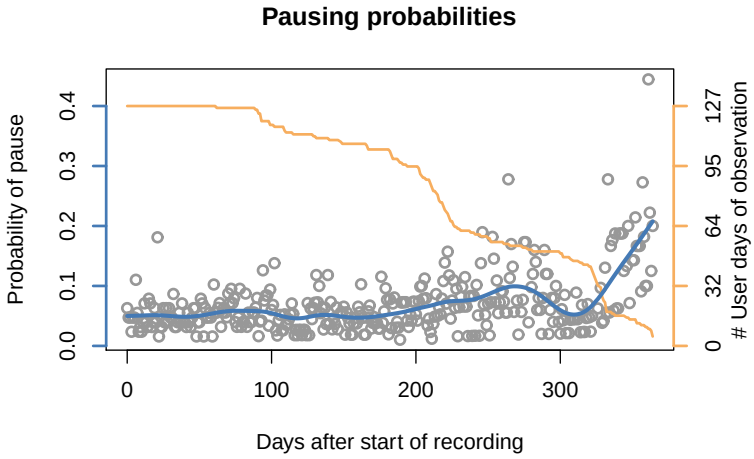


Figure S4.: Probabilities to pause GPS recording against the day of observation. In orange one can see the number of observed user days.

mast. We can see a clear bimodal behaviour, reflecting the expected division. The LTE masts are a bit sparser than those of the other two connection types, but this does not change the overall picture by much. Any method that uses proximities between cell locations should therefore be robust against differences in densities of about an order of magnitude.

3. Quality of CDR data viewed as trajectory

In the paper we only presented a table with averages. While this is sufficient to make the point that CDR data should not in general be viewed as a movement trajectory in the classical sense, we think it is nice to have a closer look at the pairwise behaviour of the individual statistics.

Figure S6 demonstrates that the correlation between the different radii of gyration is fairly high between the ones calculated on handover data and GPS. In addition their values do not depend on the amount of information we have for those days, which confirms expectations. Note that the GSM values are more dispersed, as the individual positioning of the cell centroids can influence the result. The two measures based on CDR clearly are less correlated with GPS measurements.

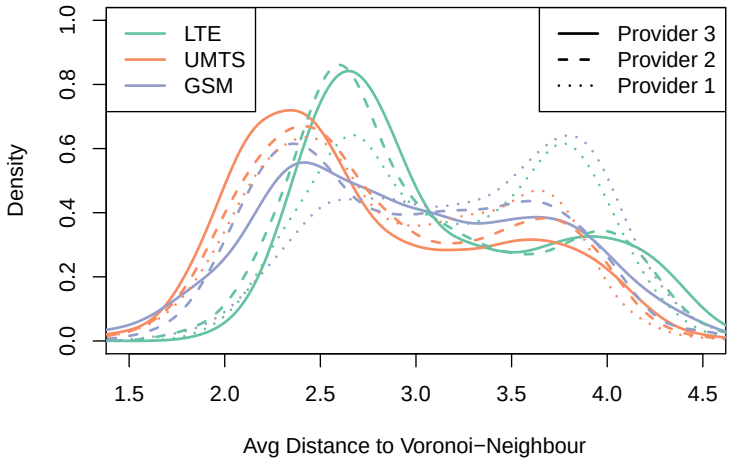


Figure S5.: Density of the average distance to Voronoi neighbours by Provider and Connection type.

Clearly visible are the days with only one CDR, typically belonging to the worst quartile in either measure of information quality.

The other statistic that was calculated were the daily distances travelled. The corresponding pairs plot is Figure S7. The GSM distance is mostly higher than the GPS distance because of handovers while the phone does not move as well as the fact that the fixes of the trajectories are the cell centres instead of the actual locations. There is an antagonising effect that movement while connected to the same cell is underreported, but this clearly is of much lower importance than the aforementioned effects. This holds to an even stronger degree if the comparison is between the GSM distances and those between GPS stops, as the latter distances are smaller than the true GPS movements. In the same vein the CDR distances are naturally lower than those based on GSM, as the fixes used in the former data source are a subset of the fixes of the latter. Interestingly the distances between GPS stops are relatively close to the distances from CDR, at least for the days with more CDR information, but naturally still very big on days with few CDR.

There seems to be a tendency for the CDR based measures to get closer to the GSM based numbers as the information increases, which is also what one would expect: In the extreme case of CDRs every second, the two measures should coincide.

In order to test this hypothesis, we plotted the logarithm to the base ten of the quotient between the distance and RoG statistics against our two quality measures in Figures S8 and S9. While there is an obvious trend towards poorer quality of the statistics with decreasing numbers of CDRs, the noise around that signal is significant and depends on the underlying true movement (for a day spent at only one location, one CDR is enough). Note also that the error is significantly larger for the distance measure, as can also be found in the literature Schulz et al. (2012). The same holds true for the noise around the error.

While the RoG based on CDR can be close to the one based on GSM data for high enough numbers of CDR, the same does not hold true of the distances. Every missed cell reduces the distance, and as the number of connected cells during a day is typically large, there is simply no hope of getting even close to the GSM distances

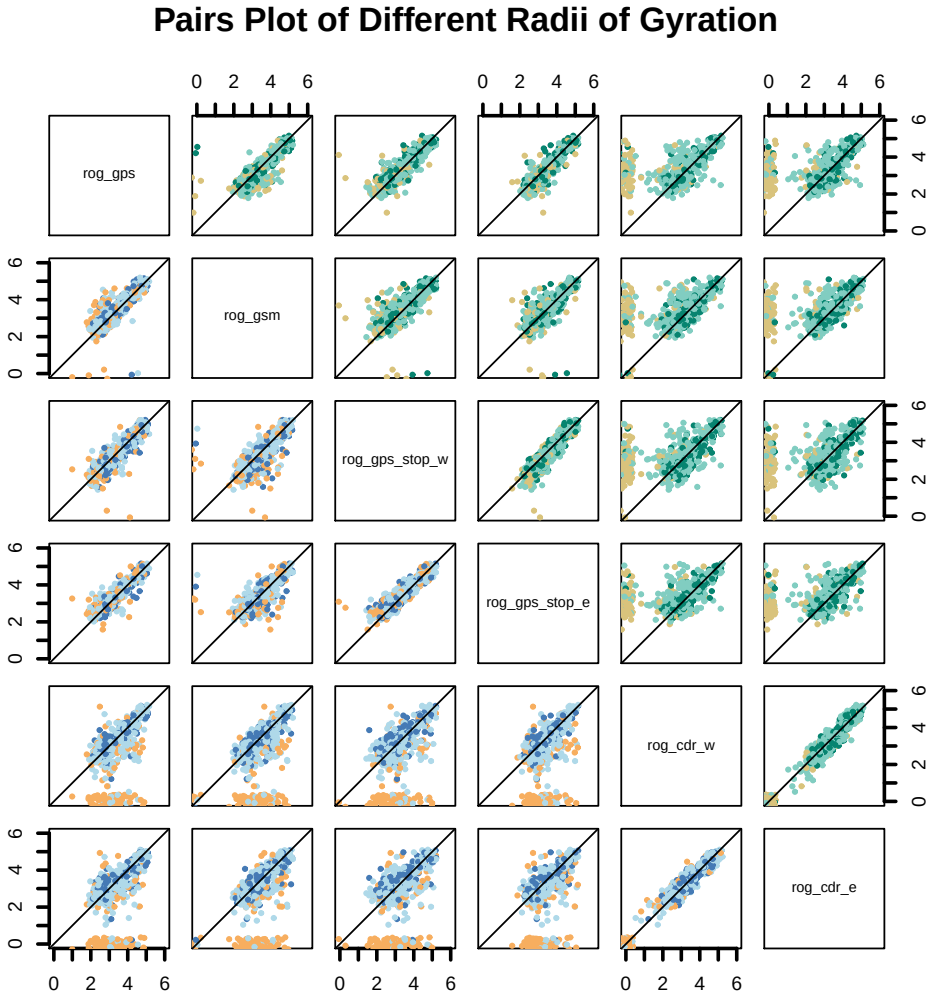


Figure S6.: Pairs plot of radii of gyration, measured in different ways. W stands for weighted RoGs, while S indicates RoGs based on the stay points of the GPS-trajectory. The points represent 600 randomly selected days from our database for which we had at least two stop points. The pairs plot is symmetrical in the positioning of points, yet not in colouring. Orange indicates the quartile with the least information, dark colours indicate the respective quartiles with the most information. Below the diagonal the amount of information is measured as the time-weighted average fraction of the day that a CDR is closest; above the diagonal, information is measured in CDR counts for that day. Some jitter was added to the locations to reveal areas of high density.

with the numbers of CDRs that we have. Interestingly however the quotient of the distance measures does not deteriorate with lower values of CDR if the distances based on GPS are in the denominator, as illustrated in Figure S10.

The figures also show the differences of typical time per CDR based on the two ways of calculating it. The fact that the points in orange are to the left of the points in blue is reminiscent of the bursty nature of CDR. Average fraction of the day covered by a CDR calculated as one over the number of CDRs decreases inversely proportional to that number. This leads to an underestimation of the

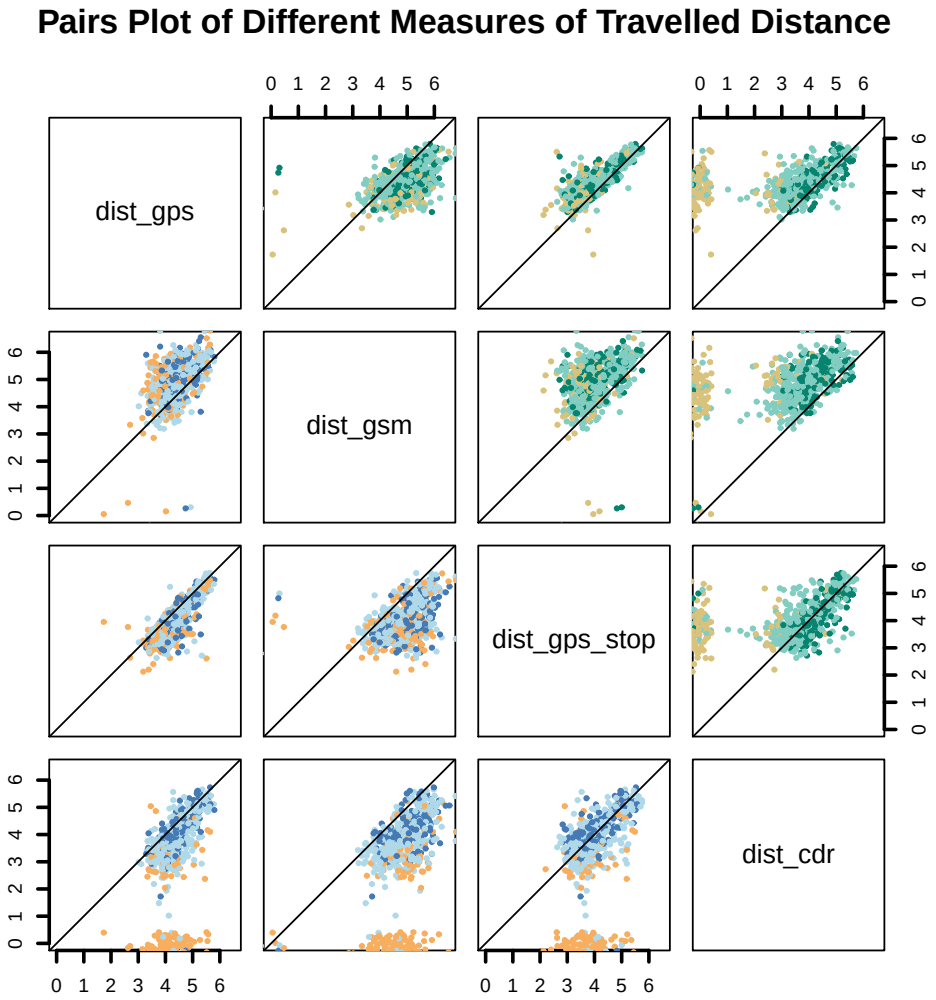


Figure S7.: Pairs plot of daily travelled distances, measured in different ways. S indicates distances based on the stay points of the GPS-trajectory. The points represent 600 randomly selected days from our database for which we had at least two stop points. The pairs plot is symmetrical in the positioning of points, yet not in colouring. Orange indicates the quartile with the least information, dark colours indicate the respective quartiles with the most information. Below the diagonal the amount of information is measured as the time-weighted average fraction of the day that a CDR is closest; above the diagonal, information is measured in CDR counts for that day. Some jitter was added to the locations to reveal areas of high density.

time for which a CDR is the closest, as seen from the Figures S9 and S8. A natural lower bound of actual (weighted) average fractions of a day covered by a CDR is the square of the fraction of the day spent sleeping (corresponding to zero time between CDRs while awake). For a sleep duration of eight hours this corresponds to 0.11, which is about the minimum of what can be observed on the actual values in blue. While passive (incoming) CDRs can be received during sleep, the bound is nonetheless an interesting benchmark.

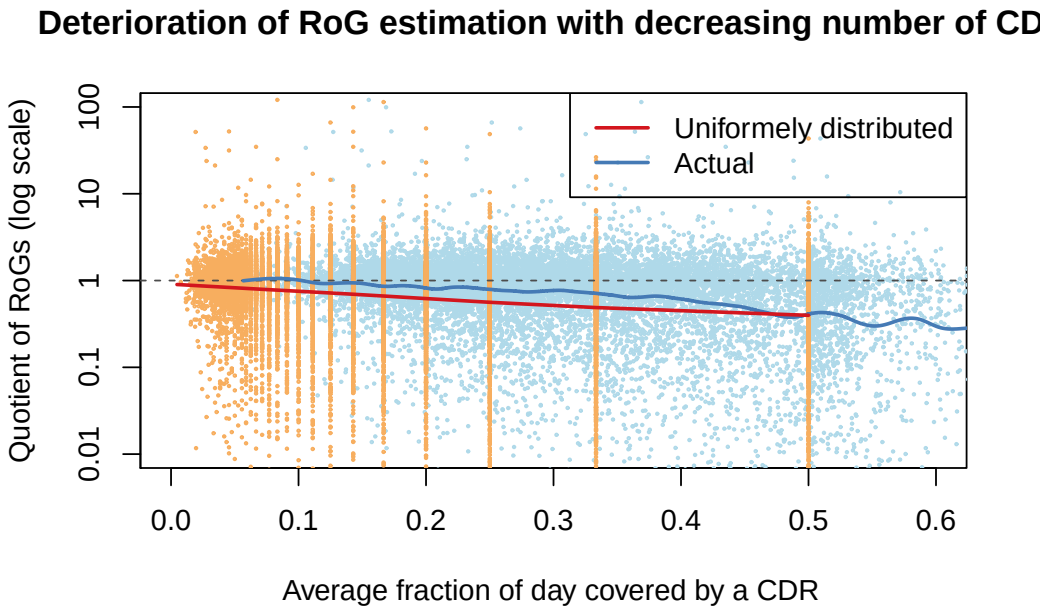


Figure S8.: Quotient of RoG calculated from the CDR-"trajectory" and the RoGs based on GSM. In blue with average observed fractions of the day covered by a CDR of that day, in orange with one over the number of CDRs that day as the average fraction. Smoothing splines for both solutions (on the log scale) in darker colors.

4. Choice of clustering

In our data we observe behaviour whose clustering is not straight forward. The reasons are similar clusters (e.g. a base cluster and a variant with a stop at a pub after work) and different calling behaviour in different routines (a person might call less in her summer cottage, leading to a larger distance). This is visualised in S11. Assuming normal work days (without pub visits) are all similar (dense cluster on the left side). The variants with bar are a bit different and a bit more dispersed (timing might be somewhat different and the pubs can vary as well). A Third hypothetical cluster is characterised by the weekend behaviour with fewer calls (and hence a larger distance to other points in the cluster). Choosing any fixed ϵ can at most separate one cluster from the other two and therefore is not sufficient. Sequential clustering with increasing ϵ first finds the dense clusters followed by the later, less dense ones.

5. Some user examples

In this section we would like to exemplify the variety of users we had by showing some plots of their data.

First we show some distance matrices. In Figure S12 one can see the behaviour of changing the number of time slots through the day. More days get included as the time slots get partitioned into finer pieces, allowing CDRs that formerly were together in one slot to stand alone. The overall appearance of the matrix does usually not change much, so the clusters should be rather stable. The number of

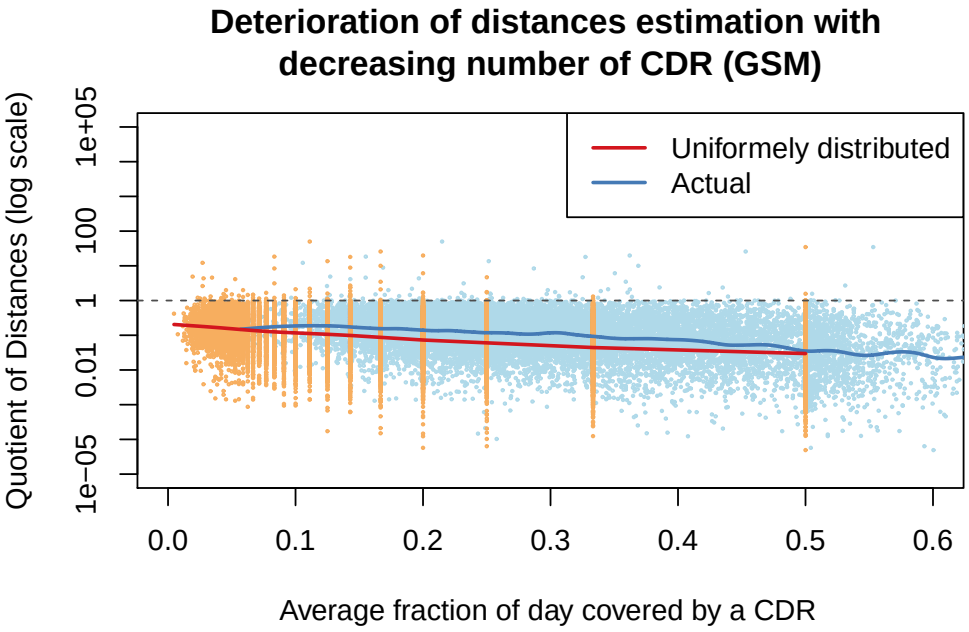


Figure S9.: Quotient of daily distances calculated from the CDR-"trajectory" and the distances based on GSM. In blue with average observed fractions of the day covered by a CDR of that day, in orange with one over the number of CDRs that day as the average fraction. Smoothing splines for both solutions (on the log scale) in darker colors.

days that are considered can only increase as nt increases, as CDR can fall into seperate time slots with finer daily segmentations. As long as there is enough data, high nt should be preferred, as they allow for finer grained statements.

In Figure S13 we show the effect of different counts of days with enough CDR to allow clustering. The image corresponding to User 251 exemplifies the problem of users with very few CDR. As the days where not enough CDRs are observed to even potentially be clustered given the chosen range ϵ are excluded from the clustering, users with few days of enough CDRs can easily be spotted by small distance matrices. The distance matrix for user 143 finally shows some days towards the end of the observation period that are not similar to a significant number of other days. Those days will typically not be matched to any cluster and for those days little can be done in the current setting.

Next we have a look at the corresponding GPS movements in Figure S14. Clealy the regime change of User 174 is visible where the main daytime location changes. The regularity of the morning calling behaviour at this new location is unique among all the users in our dataset. Nonetheless it is a good example of the method not being able to capture movement without CDR: Note that on no day there is a CDR before the green activity. Therefore this cannot be captured and consequently the whole morning is attributed to the green location, increasing the average error. User 251 produced unusually few CDR which is the reason for the low dimension of the distance matrix. For User 143 it is perhaps only with some difficulty possible to see why some days towards the end do not resemble any other days. The reason lies in the yellow locations of considerable duration that happen at different times, so the days are not deemed similar to other days. User 154 was simply added to show how irregular user behaviour can be.

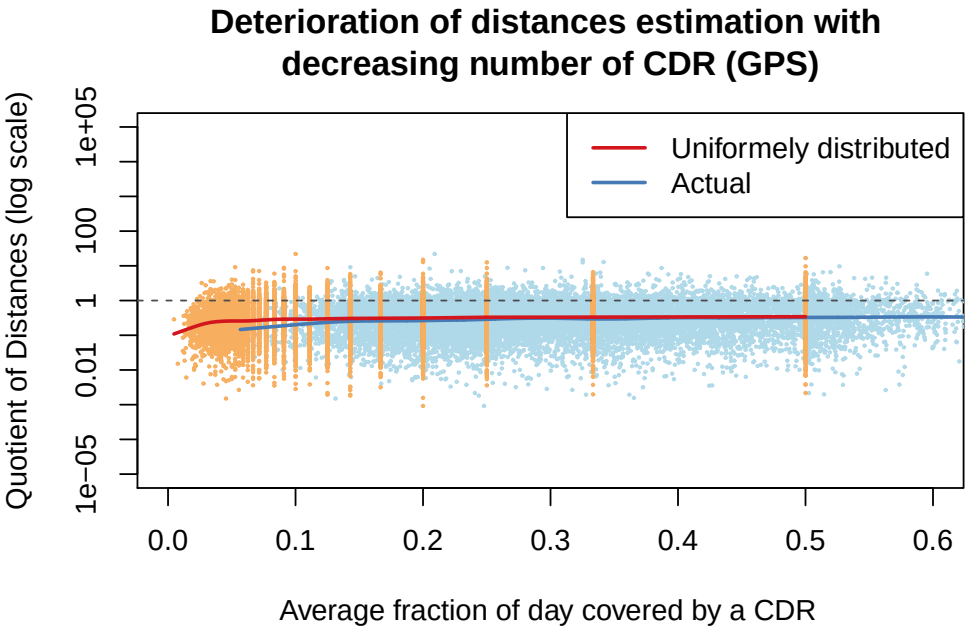


Figure S10.: Quotient of daily distances calculated from the CDR-"trajectory" and the distances based on GPS. In blue with average observed fractions of the day covered by a CDR of that day, in orange with one over the number of CDRs that day as the average fraction. Smoothing splines for both solutions (on the log scale) in darker colors.

Next we would like to show some examples of the clustering of the masts in Figure S15. Identifying clusters that are far apart, as in the large scale panel is easy. It is within cities, as shown in the small scale panel where the tuning of the parameters poses a challenge. If the clusters are chosen too small, identifying similar days becomes more difficult, as the clusters of the same GPS point do not have to bear the same label. Choosing them too big will yield trivial results for people whose important locations are relatively close together. Note that not all significant GPS stops are plotted to avoid overcrowding.

Last, we would like to show some reconstructions of daily movements. In Figure S16 we have plotted the reconstruction (in green) against GPS (red) and handover (black) data. In Figure 16(a) we see a nice home–work–home pattern that is well captured by the reconstruction, even if the precise timing of coming back home is somewhat off. In the distance between the red on the one hand and the black and green trajectories on the other hand we see that the true (GPS) location differs from the GSM location due to the spatial granularity of the mast locations. Also we see that the non-GPS signal usually do not capture the exact route that was chosen, as information between stops is often times missing. In Figure 16(b), we show an unusual day for user 58. In particular, there was no matching pattern that included the location visited that day. As a consequence, the missing time slots are assumed to be the last known location, which in this instance differs from the information from the GPS trajectory quite substantially.

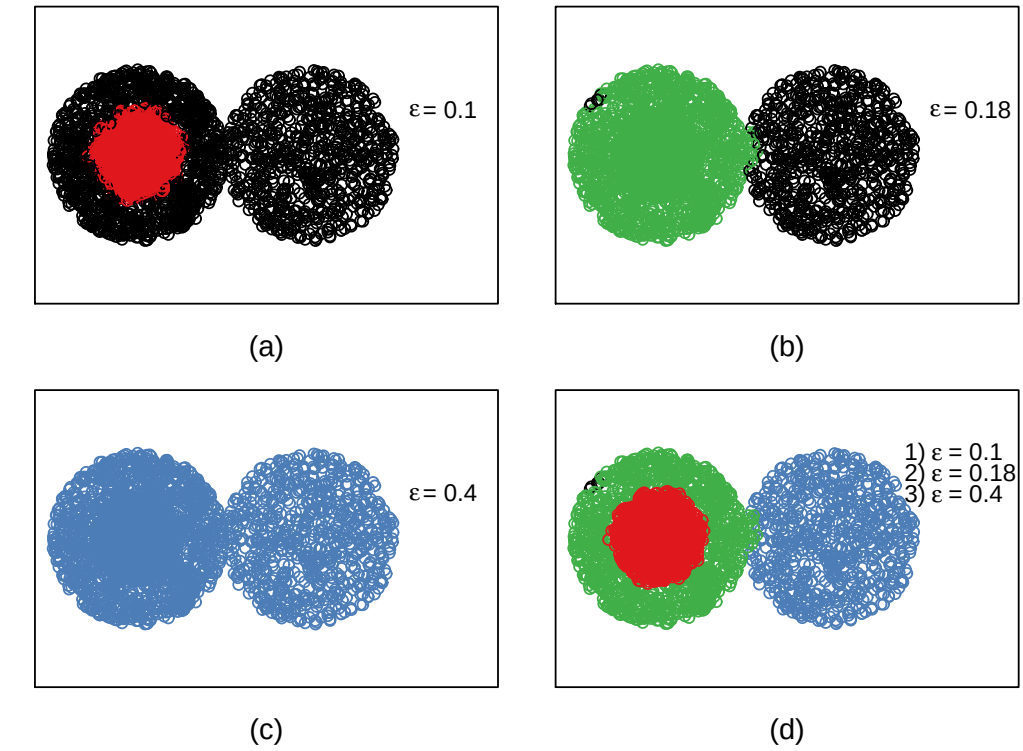


Figure S11.: Clustering three clusters in the following setup: Two clusters have the same centre but one is a bit more dispersed; the third is further away and even more dispersed. Choosing a fixed ϵ cannot separate the three, but a sequential clustering with increasing ϵ can.

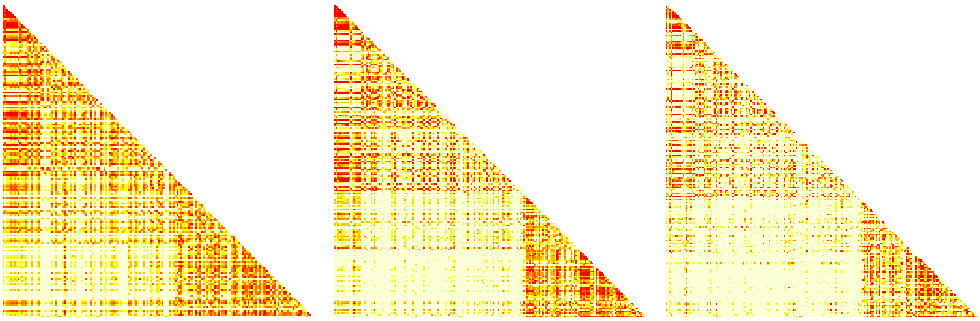
References

John Steenbruggen, Emmanouil Tranos, and Peter Nijkamp. Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy*, 39(3-4):335–346, 2015. .

Salvatore Rinzivillo, Simone Mainardi, Fabio Pezzoni, Michele Coscia, Dino Pedreschi, and Fosca Giannotti. Discovering the Geographical Borders of Human Mobility. *KI Künstliche Intelligenz*, 26(3):253–260, 2012. .

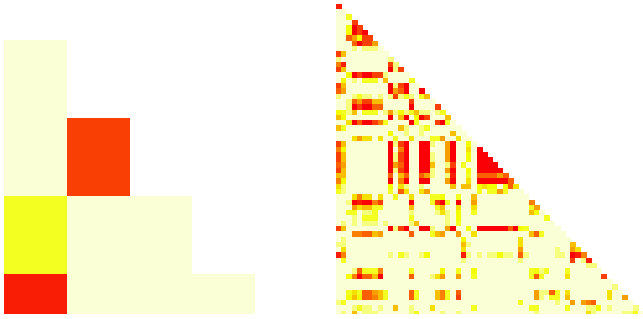
Kevin S. Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE*, 9(6), 2014. .

Daniel Schulz, Sebastian Bothe, and Christine Körner. Human mobility from gsm data-a valid alternative to gps? In *Mobile data challenge 2012 workshop, June*, pages 18–19, 2012.



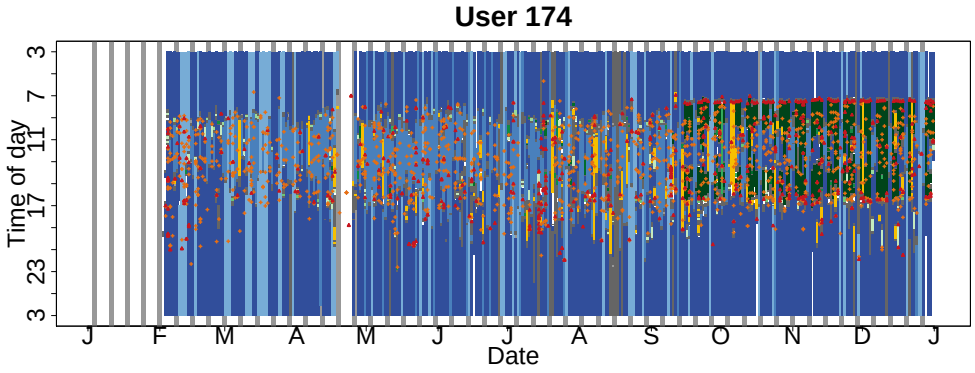
(a) $nt = 6$ (b) $nt = 12$ (c) $nt = 24$

Figure S12.: Distance matrices between days for User 174 for $nt = 6$, $nt = 12$ and $nt = 24$. They are qualitatively the same, but there are fewer days available for $nt = 6$.

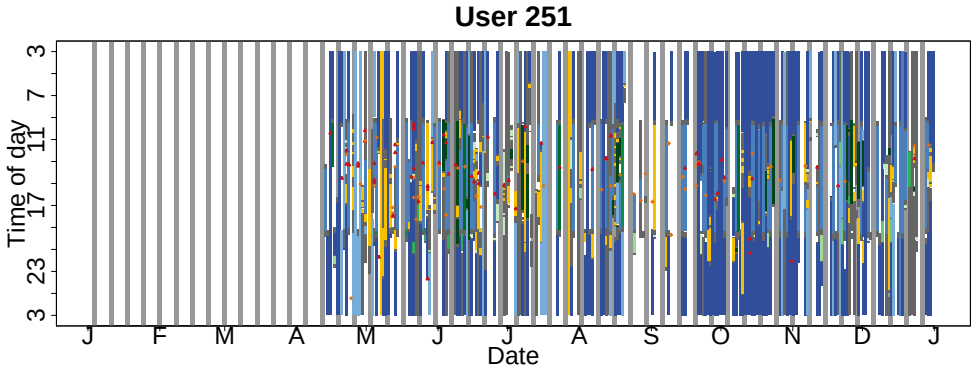


(a) User 251 (b) User 143

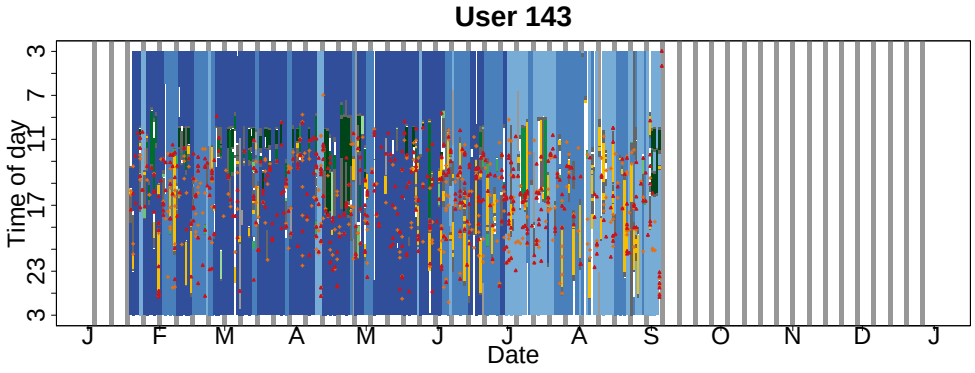
Figure S13.: Distance matrices for users with different number of days with at least the minimum CDR count for clustering. User 251 has very few days with enough CDR to fill the minimal requirement for clustering, User 143 has more.



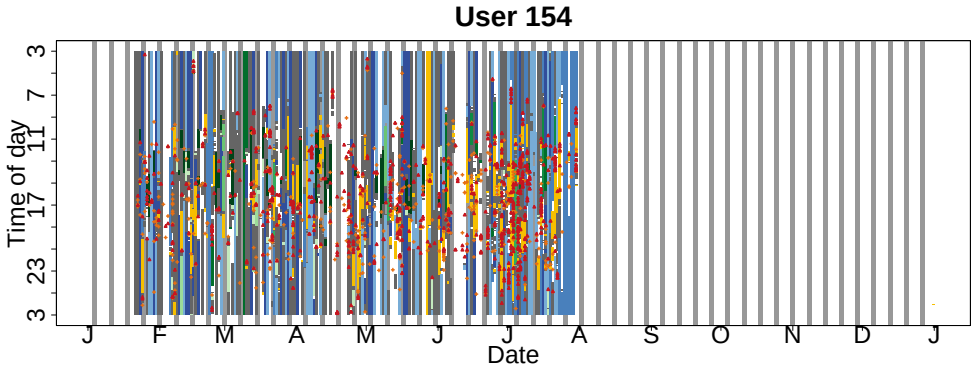
(a) User 174



(b) User 251

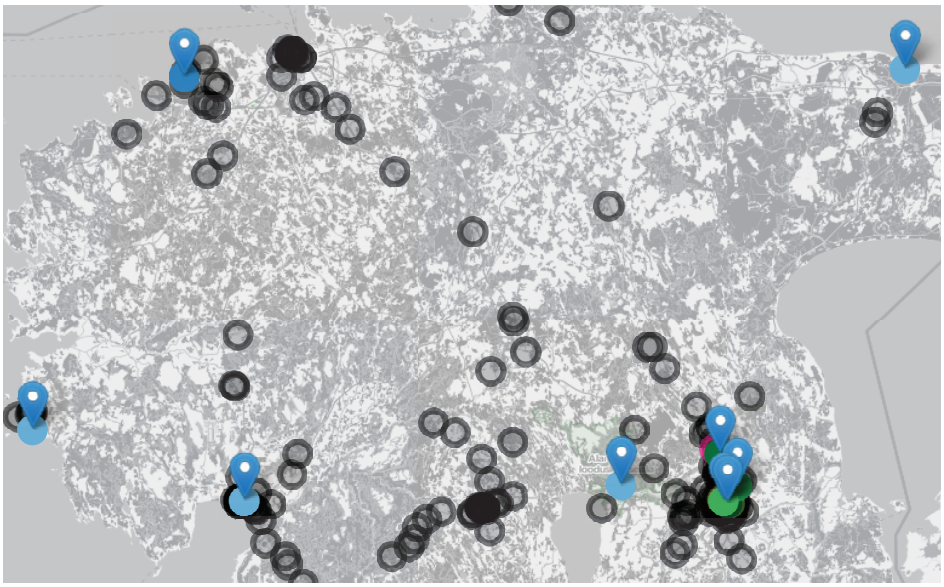


(c) User 143

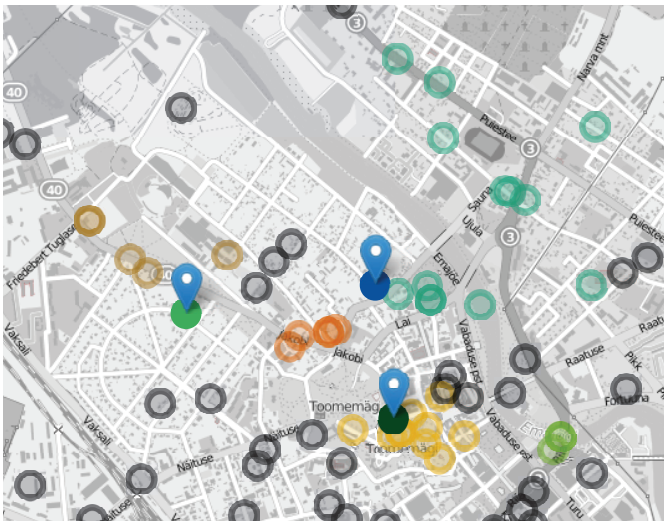


(d) User 154

Figure S14.: Visualisations of the GPS signals and the CDR for different users over the whole study period.

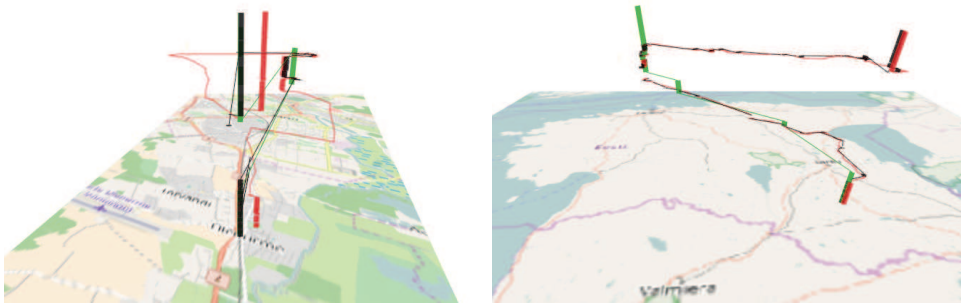


(a) Large scale



(b) Small scale

Figure S15.: Examples of the mast clustering for a user. Unclustered masts are shown in gray, clustered masts are in color. Markers point to locations that were found important from the GPS signal. Note the different spatial extent of the clusters in the small scale image that is a result of the adjusted distances.



(a) User 37 (b) User 58

Figure S16.: GPS (red), handover (black) and reconstructed (green) trajectories of one day each of two users. Clearly the reconstruction can at most capture what is in the handover data and only if there is enough repetition.